digital humanities
Lausanne – Switzerland '14

UNIL | Université de Lausanne

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

COMBINING MACRO- AND MICROANALYSIS FOR EXPLORING THE CONSTRUAL OF SCIENTIFIC DISCIPLINARITY

Fankhauser, Peter
IDS Mannheim, Germany

Kermes, Hannah
Saarland University, Germany

Teich, Elke
Saarland University, Germany

Category: Poster
Session: 2
Date: 2014-07-10
Time: 16:00:00
Room: Amphipôle Common Area

1. Introduction

The English Scientific Text Corpus (SciTex) consists of about 5000 scientific papers with about 34 Mio tokens in two time slots, 1970/80s and 2000s [1], [2]. It has been compiled to investigate the construal of scientific disciplinarity, in particular, how interdisciplinary *contact* disciplines emerge from their *seed* disciplines. Both time slots consist of nine disciplines: Computer Science (A) as one seed discipline, Linguistics (C1), Biology (C2), Mechanical Engineering (C3), Electrical Engineering

(C4) as the other seed disciplines, and Computational Linguistics (B1), Bioinformatics (B2), Digital Construction (B3), and Microelectronics (B4) as the corresponding contact disciplines between A and C1-C4. The individual articles are subdivided into Abstract, Introduction, Main, and Conclusion.

The orthogonal dimensions time, discipline, and logical structure provide for many, potentially interesting setups of variational analysis: We can explore the diachronic evolution of contact disciplines in comparison to their seed disciplines, variation between contact disciplines and their seed disciplines, and genre variation between abstracts and text bodies in individual disciplines. In this paper we present an approach that combines a macroanalytic perspective [3] with the more traditional microanalytic perspective served by concordance search to explore variation along these dimensions.

2. Approach

2.1. Macroanalysis

For supporting explorative macroanalysis, we use well understood visualization techniques – heatmaps and wordclouds – and combine them with intuitive exploration paradigms – drill down and side by side comparison (see Figure 1). The heatmaps and wordclouds are interactive, allowing for a closer inspection at various levels. The leftmost heatmap visualizes the highest level contrast between abstracts and text bodies in the two time slots (1970s/80s and 2000s). The middle and right heatmaps serve for inspecting a chosen contrast at a lower level at the level of individual disciplines. A particular contrast can be chosen by clicking on the respective square, numbers indicating which contrast is displayed in the middle (Selection 1) and right heatmap (Selection 2). In this example, the middle heatmap visualizes the distances between abstracts and text bodies, and the right heatmap visualizes the distances between text bodies and abstracts.

The wordclouds underneath the heatmaps display the most typical words for a chosen contrast. In Figure 1 the wordcloud to the left visualizes the most typical words for abstracts as opposed to text bodies in the 2000s. Unlike in the common use of wordclouds, the size of words is proportional to their contribution to the distance (as defined in Section 2.2), whereas relative frequency is visualized by color, ranging from purple to red.

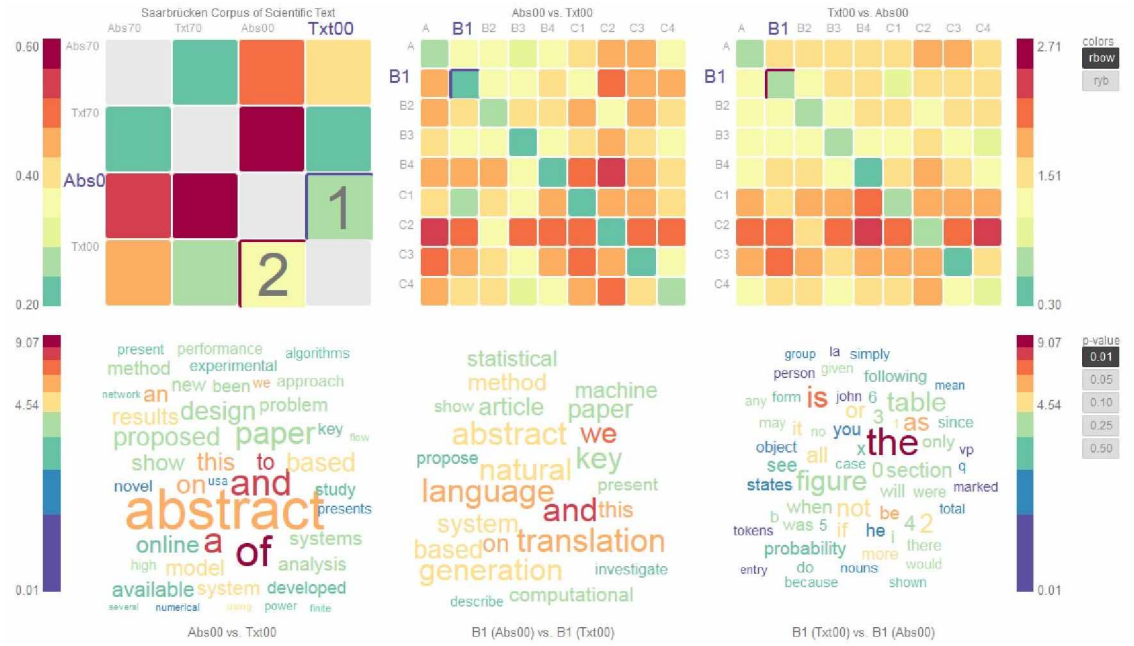


Fig. 1: Contrast between Abstracts and Text Bodies

Having a closer look at Figure 1, we can observe that the distance between abstracts is generally larger than the distance between text bodies, and that it has increased in the 30 years period. This general trend is mirrored in the individual disciplines (not shown here). Looking at the middle and right heatmaps, we can see that - not surprisingly - the distance between particular disciplines are at a minimum (squares forming the main diagonal), and the distances among the seed disciplines (A and C corpora), are generally larger than the distances among contract disciplines.

The corresponding wordclouds visualize the most typical words for abstracts (middle heatmap) and for text bodies (right heatmap) in the discipline B1 (Computational Linguistics). In this particular contrast, words typical for abstracts are clearly centered around constructions of exposition (*we propose, describe, investigate*), main topics of B1 (*natural, language (generation), machine translation*), words describing the methodology (*method, statistical, computational, system*) and function words (*and, of, on*). Words typical for text bodies are markedly different: they comprise B1's main entities of topic elaboration (*tokens, nouns, object, vp, john, probability*), references (*see figure, table, section*), conjunctions (*when, since, because, if*), auxiliary and modal verbs (*be, is, was, were, do, will, would, may*), and prominently, the determiner *the*. In summary, abstracts exhibit characteristics of an

informationally dense text (e.g., omission of determiners) with topic oriented content. In contrast, text bodies are less dense (determiners, references, modality) and more elaborated.

Other contrastive pairs, such as the synchronic comparison between disciplines or the diachronic comparison of the two time slots, corroborate the results derived by means of computationally much more demanding techniques from machine learning [1], [2].

2.2. Corpus Representation and Distance Measures

The individual corpora are tokenized, and tokens are transformed to lower case. Stopwords are deliberately not excluded to inspect all levels of variation: style, lexico-grammar, and theme. On this basis, corpora are represented by means of unigram language models smoothed with Jelinek-Mercer smoothing, which is a linear interpolation between the relative frequency of a word in a subcorpus and its relative frequency in the entire corpus [4]. The distance between two corpora P and Q is measured by relative entropy D, also known as Kullback-Leibler Divergence:

$D(P||Q)>=\text{Sum\_}w\ p(w)*\log\_2(p(w)/q(w))$

Here $p(w)$ is the probability of a word w in P, and $q(w)$ is its probability in Q. Relative entropy thus measures the average amount of *additional* bits per word needed to encode words distributed according to P by using an encoding optimized for Q. Note that this measure is asymmetric, i.e., $D(P||Q)$!= $D(Q||P)$, and has its minimum at 0 for $P = Q$[5].

The individual word weights are calculated by the pointwise Kullback-Leibler Divergence [6]:

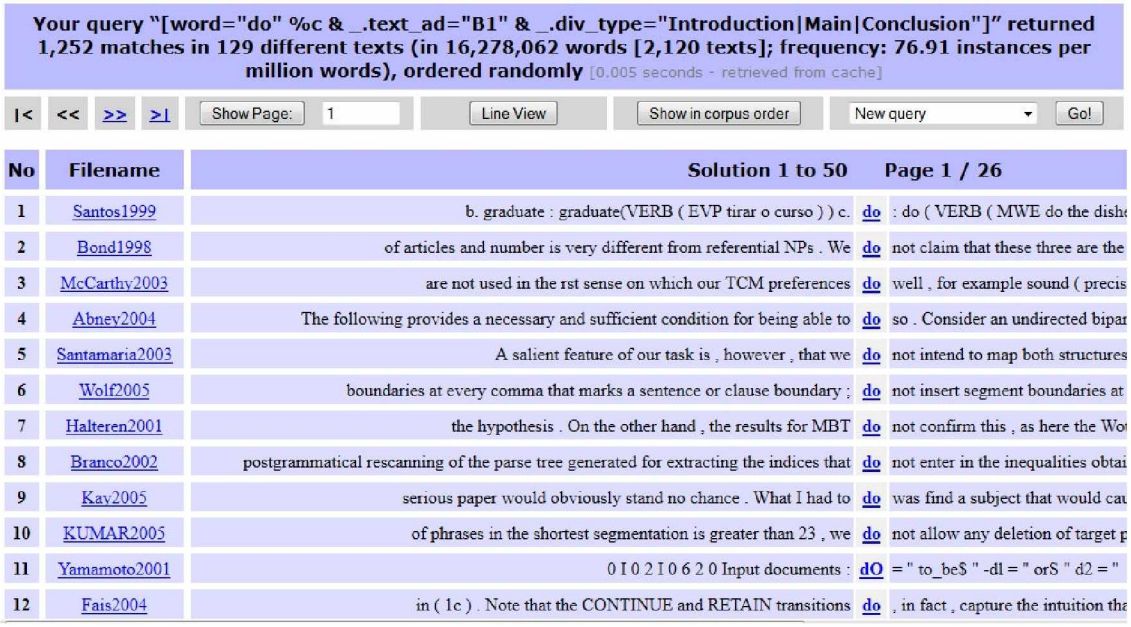
$D\_w(P||Q) = p(w)*\log\_2(p(w)/q(w))$

For all words the statistical significance of a difference is calculated based on an unpaired Welch t-test on the observed word probabilities in the individual documents of a corpus. This is used for discarding words below a given level of significance (p-value). A more detailed comparison with other measures for comparing corpora [7] is beyond the scope of this paper and will appear in another venue.

2.3. Microanalysis

Wordclouds serve as a bridge between the big distance picture of macroanalysis and microanalyis. To this end, they are seamlessly integrated with the IMS Open Corpus Workbench (CQPWeb: http://cwb.sourceforge.net/index.php), which provides for an expressive corpus query language and

several summarization tools, such as collocations and comparative word frequency lists. A click on a word sends a query to CQPWeb, which returns the word in the chosen context. For example, clicking on “do” in the right heatmap (B1 (Txt00) vs. B1 (Abs00)) generates the following query shown in Figure 2.



Your query “[word="do" %c & _.text_ad="B1" & _.div_type="Introduction|Main|Conclusion"]” returned 1,252 matches in 129 different texts (in 16,278,062 words [2,120 texts]; frequency: 76.91 instances per million words), ordered randomly [0.005 seconds - retrieved from cache]

|< << >> >1 Show Page: 1 Line View Show in corpus order New query Go!

No	Filename	Solution 1 to 50	Page 1 / 26		
1	Santos1999	b. graduate : graduate(VERB (EVP tirar o curso)) c.	do	: do (VERB (MWE do the dish
2	Bond1998	of articles and number is very different from referential NPs . We	do	not claim that these three are the
3	McCarthy2003	are not used in the rst sense on which our TCM preferences	do	well , for example sound (precis
4	Abney2004	The following provides a necessary and sufficient condition for being able to	do	so . Consider an undirected bipar
5	Santamaria2003	A salient feature of our task is , however , that we	do	not intend to map both structures
6	Wolf2005	boundaries at every comma that marks a sentence or clause boundary ;	do	not insert segment boundaries at
7	Halteren2001	the hypothesis . On the other hand , the results for MBT	do	not confirm this , as here the Wo
8	Branco2002	postgrammatical rescanning of the parse tree generated for extracting the indices that	do	not enter in the inequalities obtai
9	Kay2005	serious paper would obviously stand no chance . What I had to	do	was find a subject that would cau
10	KUMAR2005	of phrases in the shortest segmentation is greater than 23 , we	do	not allow any deletion of target p
11	Yamamoto2001	0 I 0 2 I 0 6 2 0 Input documents :	dO	= " to_be$ " -dl = " orS " d2 = "
12	Fais2004	in (1c) . Note that the CONTINUE and RETAIN transitions	do	, in fact , capture the intuition tha

Fig. 2: Concordance for “do” in B1, text bodies, 2000s

This query returns a concordance for “do” in the 2000s slot of SciTex constrained to subcorpus B1 and to the divisions Introduction, Main, and Conclusion. Based on this list one can inspect the larger context of individual hits and get a ranked list of collocations to distinguish the uses of “do” as an auxiliary vs. main verb.

3. Related Work

The need for combining macroanalysis with microanalysis is well recognized in the DH community [8], [9], and there does exist a variety of frameworks with similar goals. Due to space restrictions, we can only give an exemplary selection; for a comprehensive overview see TAPoR 2.0 (http://tapor.ca/). The MONK workbench [10] allows to compare pairs of corpora using Dunning's log-likelihood ratio [11] for word weighting. Apart from the different distance measure, the main difference of our approach is that

we combine the macro perspective of overall distance with the micro perspective of individual word weights to allow for an explorative analysis of variation. The Voyant Tools [12] provide a plethora of text visualizations, including word clouds, cooccurrences, and word trends based on frequencies. The focus of these tools, however, lies on summarizing and visualizing one text or corpus, rather than on exploring variation among corpora. Finally, the TXM platform [13] integrates the IMS Corpus Workbench with some macroanalysis R packages such as factorial correspondence analysis, contrastive word specificity, and cooccurrence analysis. While this integration certainly provides a broader set of analysis techniques, it is arguably more complicated to use than the system presented in this paper.

4. Summary and Future Work

We have presented a system that combines macroanalysis with microanalysis to explore language variation, and briefly illustrated its use for analyzing differences along the dimensions time, discipline, and genre in a corpus of scientific text. Future work will be devoted both to technical as well as methodological enhancements. A useful technical extension is the facility to interactively group subcorpora to larger units, maybe with the help of hierarchical clustering based on the distance matrix to form meaningful groups. More generally, the support for importing external corpora and exporting distance matrices and word weights for analysis with other tools is desirable – the presented system has been evaluated based on a number of corpora, but the underlying processing pipeline certainly needs to be generalized and improved. On the methodological side the main challenge lies in supporting a broader variety of feature sets beyond simple unigram language models. This includes latent language models such as topic models [14] and hidden markov models [15], but also enriched representations such as part-of-speech tagging, and other extensions of unigram models. Such richer feature sets allow to focus analysis by means of feature selection, but also bear new challenges in measuring and visualizing the contribution of features to a contrast at hand, and translating features into meaningful queries against the underlying corpus.

References

1. **Elke Teich and Peter Fankhauser** (2010). *Exploring a Corpus of Scientific Texts using Data Mining*. In S. Gries, S. Wulff, and M. Davies, editors, Corpus-linguistic applications: Current studies,

new directions, pp. 233–247. Rodopi, Amsterdam and New York.

2. **Stefania Degaetano-Ortlieb, Hannah Kermes, Ekaterina Lapshinova-Koltunski, and Elke Teich** (2013). *SciTex: A diachronic corpus for analyzing the development of scientific registers*. In Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt, editors, New Methods in Historical Corpus Linguistics, Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP), Volume 3, Narr, Tübingen.

3. **Matthew L. Jockers** (2013). *Macroanalysis: Digital Methods* & Literary History. University of Illinois Press, Urbana, Chicago, and Springfield.

4. **Chengxiang Zhai and John Lafferty** (2004). *A study of smoothing methods for language models applied to information retrieval*. ACM Transactions on Information Systems (TOIS), 22(2):179–214.

5. **David J. C. MacKay** (2002). *Information Theory*, Inference & Learning Algorithms. Cambridge University Press, New York, NY, USA.

6. **Takashi Tomokiyo and Matthew Hurst** (2003). *A language model approach to keyphrase extraction*. Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment (MWE '03), Vol. 18, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 33–40. DOI=10.3115/1119282.1119287 dx.doi.org/10.3115/1119282.1119287

7. **Adam Kilgarriff** (2001). *Comparing Corpora*. International Journal of Corpus Linguistics, 6(1):97–133.

8. **Michael Correll and Michael Gleicher** (2012). *What Shakespeare Taught Us About Text Visualization*. IEEE Visualization Workshop Proceedings, 2nd Workshop on Interactive Visual Text Analytics: Task-Driven Analysis of Social Media Content, Seattle, Washington, USA, Oct 2012.

9. **Matthew L. Jockers and Julia Flanders** (2013). *A Matter of Scale*. Staged debate at the Boston Area Days of Digital Humanities Conference at Northeastern University, March 18, 2013. digitalcommons.unl.edu/englishfacpubs/106/

10. **John Unsworth and Martin Mueller** (2009). *The MONK Project Final Report*. Sep 2009. www.monkproject.org/MONKProjectFinalReport.pdf

11. **Ted Dunning** (1993). *Accurate methods for the statistics of surprise and coincidence*. Computational Linguistics 19(1):61–74.

12. **Stéfan Sinclair, Geoffrey Rockwell and the Voyant Tools Team** (2012). *Voyant Tools* (web application). http://voyant-tools.org/

13. **Serge Heiden** (2010). *The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme*. Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, Institute for Digital Enhancement of Cognitive Development, Waseda University, Japan, Nov 2010, pp. 389-398.

14. **David. M. Blei, Andrew Y. Ng, and Michael I. Jordan** (2003). *Latent Dirichlet Allocation*. Journal of Machine Learning Research, 3:993–1022.

15. **Sharon Goldwater and Tom Griffiths** (2007). *A fully Bayesian approach to unsupervised part-of-speech tagging*. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07). Association for Computational Linguistics, Prague, Czech Republic, June 2007, pp. 744–751. www.aclweb.org/anthology/P07-1094